Big data and positive social change in the developing world:

A white paper for practitioners and researchers

Rockefeller Foundation Bellagio Centre conference, May 2014

oii oxford internet institute university of oxford

THE ROCKEFELLER FOUNDATION

UNIVERSITY OF AMSTERDAM

ALFRED P. SLOAN FOUNDATION

Please cite as:

Bellagio Big Data Workshop Participants. (2014). "Big data and positive social change in the developing world: A white paper for practitioners and researchers." Oxford: Oxford Internet Institute. Available online: http://ssrn.com/abstract=2491555.

Contents

Summary

This white paper was produced by a group of activists, researchers and data experts who met at the Rockefeller Foundation's Bellagio Centre to discuss the question of whether, and how, big data is becoming a resource for positive social change in low- and middle-income countries (LMICs). Our working definition of big data includes, but is not limited to, sources such as social media, mobile phone use, digitally mediated transactions, the online news media, and administrative records. It can be categorised as data that is **provided explicitly** (e.g. social media feedback); data that is **observed** (e.g. mobile phone call records); and data that is **inferred and derived by algorithms** (for example social network structure or inflation rates). We defined four main areas where big data has potential for those interested in promoting positive social change: advocating and facilitating; describing and predicting; facilitating information exchange and promoting accountability and transparency.

In terms of **advocating and facilitating**, we discussed ways in which volunteered data may help organisations to open up new public spaces for discussion and awareness-building; how both aggregating data and working across different databases can be tools for building awareness, and how the digital data commons can also configure new communities and actions (sometimes serendipitously) through data science and aggregation. Finally, we also looked at the problem of overexposure and how activists and organisations can protect themselves and hide their digital footprints. The challenges we identified in this area were how to interpret data correctly when supplementary information may be lacking; organisational capacity constraints around processing and storing data, and issues around data dissemination, i.e. the possible negative consequences of inadvertently identifying groups or individuals.

Next, we looked at the way big data can help **describe and predict**, functions which are particularly important in the academic, development and humanitarian areas of work where researchers can combine data into new dynamic, high-resolution datasets to detect new correlations and surface new questions. With data such as mobile phone data and Twitter analytics, understanding the data's comprehensiveness, meaning and bias are the main challenges, accompanied by the problem of developing new and more comprehensive ethical systems to protect data subjects where data is observed rather than volunteered.

The next group of activities discussed was **facilitating information exchange**. We looked at mobile-based information services, where it is possible for a platform created around a particular aim (e.g. agricultural knowledge-building) to incorporate multiple feedback loops which feed into both research and action. The pitfalls include the technical challenge of developing a platform which is lean yet multifaceted in terms of its uses, and particularly making it reliably available to low-income users. This kind of platform, addressed by big data analytics, also offers new insights through data discovery and allows the provider to steer service provision according to users' revealed needs and priorities.

Our last category for big data use was **accountability and transparency**, where organisations are using crowdsourcing methods to aggregate and analyse information in real time to establish new spaces for

critical discussion, awareness and action. Flows of digital information can be managed to prioritise participation and feedback, provide a safe space to engage with policy decisions and expose abuse. The main challenges are how to keep sensitive information (and informants) safe while also exposing data and making authorities accountable; how to make the work sustainable without selling data, and how to establish feedback loops so that users remain involved in the work beyond an initial posting. In the crowdsourcing context, new challenges are also arising in terms of how to verify and moderate real-time flows of information, and how to make this process itself transparent.

Finally, we also discussed **the relationship between big and open data**. Open data can be seen as a system of governance and a knowledge commons, whereas big data does not by its nature involve the idea of the commons, so we leaned toward the term '*opening* data', i.e. processes which could apply to commercially generated as much as public-sector datasets. It is also important to understand where to prioritise opening, and where this may exclude people who are not using the 'right' technologies: for example, analogue methods (e.g. nailing a local authority budget to a town hall door every month) may be more open than 'open' digital data that's available online.

Our discussion surfaced many questions to do with **representation and meaning**: must datasets be interpreted by people with local knowledge? For researchers to get access to data that is fully representative, do we need a data commons? How are data proprietors engaging with the power dynamics and inequalities in the research field, and how can civil society engage with the private sector on its own terms if data access is skewed towards elites? We also looked at issues of **privacy and risk**: do we need a contextual risk perspective rather than a single set of standards? What is the role of local knowledge in protecting data subjects, and what kinds of institutions and practices are necessary? We concluded that there is a case to be made for building a data commons for private/public data, and for setting up new and more appropriate ethical guidelines to deal with big data, since aggregating, linking and merging data present new kinds of privacy risk. In particular, organisations advocating for opening datasets must admit the limitations of anonymisation, which is currently being ascribed more power to protect data subjects than it merits in the era of big data.

Our analysis makes a strong case that **it is time for civil society groups in particular to become part of the conversation about the power of data**. These groups are the connectors between individuals and governments, corporations and governance institutions, and have the potential to promote big data analysis that is locally driven and rooted. Civil society groups are also crucially important but currently underrepresented in debates about privacy and the rights of technology users, and civil society as a whole has a responsibility for building critical awareness of the ways big data is being used to sort, categorise and intervene in LMICs by corporations, governments and other actors. Big data is shaping up to be one of the key battlefields of our era, incorporating many of the issues civil society activists worldwide have been working on for decades. We hope that this paper can inform organisations and individuals as to where their particular interests may gain traction in the debate, and what their contribution may look like.

1. Introduction

This paper was produced by a group of activists, researchers and data experts who met at the Rockefeller Foundation's Bellagio Centre to discuss the question of whether, and how, big data is becoming a resource for positive social change in low- and middle-income countries (LMICs). This paper was authored collaboratively and as much as possible reflects the perspectives of all participants. The names and details of the participants are listed in Appendix 1.

Our focus is on the uses of big data in LMICs in ways which are locally driven and locally relevant for societal processes. Numerous projects have focused on the growing importance of big data in humanitarian response and in large institutionally driven development projects, but little attention has been given so far to big data as a tool for local activism, advocacy, empowerment or projects focusing on specific marginalised or excluded groups. One reason for this is lack of access to data and information, but there are many others (see section 5). This paper therefore asks **what lessons local groups and researchers can draw from existing big data projects** and how they might apply those lessons on a national and local basis. Without assuming that big data analytics should become universal tools for social change or that they should supplant existing strategies, we ask how national civil society groups can become more data-aware, how their concerns might benefit from insights being produced by international networks, and **under what circumstances big data might become an important tool for smaller-scale groups working for social change**.

We consider the usefulness of big data to a broad range of communities engaged in pushing for positive social change. These include activists on digital issues such as personal data protection and privacy; organisations working on accountability and transparency; funders and international agencies aiming to promote social change and interested in using new tools and resources; researchers and policymakers working on economic or human development for whom digital data is a central resource; and researchers working with big data to inform development or humanitarian action.

We aim to provide practical information on what works and what does not – and how to solve problems – based on case studies of recent projects. We draw lessons from these real-life applications of data science which we hope will be useful to organisations and authorities to judge how data should flow and to whom, how to protect people who use digital technologies from the misuse of their data and potential related harms, and how to encourage new uses of large-scale digital data in civil society.

Some of the central questions that arose in our discussions were broader and relate to the way data helps us to understand and portray social issues such as inequality and representation. The inevitable categorisation that accompanies the analysis of digital data tends to 'flatten out' some of the differences and nuances that are important in understanding the historical and structural aspects of inequality such as gender, ethnicity, economic or social status and religion. The way a database is formed will inevitably emphasise certain characteristics over others. These choices are often made for historical reasons – France and Rwanda, for instance, do not collect statistics on religion or ethnicity. Paradoxically, this effort not to profile may lead to problems of representation by making racial and ethnic violence, for example, invisible. In connection with this question of representation and profiling, we also considered the power of the data analyst, and the responsibility to consider whether everything should be

measured just because measurement is possible. Information expressed as 'Data' makes social categories more salient so that the data scientist may cement differences which were previously more fluid.

Another set of persistent questions were about the power relationships involved in the tools and processes involved in collecting and processing data. Are we in an inexorable loop of data maximisation, where 'datafication' only generates demands for more and more data? And in connection with this, the important question of auditing data – understanding what is being emitted, where it is flowing and what it is being used for. We question whether the tools exist yet to follow our personal data, or the data it generates about us when analysed. This constant evolution of data processing technologies and practices makes it hard to know exactly who the players are in terms of power over data, and even harder to establish standards and principles for the safe and positive use of big data.

In this report we first provide some background on how we defined big data for our purposes. We then discuss the four main categories of use that formed the primary focus of our discussion: advocating and facilitating; describing and predicting; facilitating information exchange, and accountability and transparency. For each we look at the main characteristics of that category, provide some examples to illustrate current uses and projects, and analyse the issues they have raised and how the groups involved have responded to them. Next, we outline our discussion of a larger debate about 'big' versus 'open' data, and how the intersection of the two is an important question for civil society groups aiming to work with big data. Finally, we offer some conclusions on the discussion as a whole, including a list of the most important obstacles we identified along the way and our suggestions for mitigating them.

2. Background: how should we define big data in the context of working towards social change?

Big data terminology arose more than a decade ago from the corporate sphere,[1] but needed some rethinking to fit with the discussion on how this type of data might be useful to activists and civil society more generally. Therefore, although the corporate world's definition of big data has remained fairly stable around the idea of the 'three V's' (Volume, Velocity and Variety), for the purposes of our discussion we took a broader perspective which includes the idea of *datafication* – the shift toward digital production and dissemination of information which has traditionally come in analogue form, for example media and statistical records. Our perspective on big data also incorporates problems of interpretation (a 'fourth V' – Veracity). Our working definition of big data therefore refers to **digital datasets of unprecedented size in relation to a particular question or phenomenon**, and particularly **datasets that can be linked, merged and analysed in combination**. Practically, we suggest that it may be more relevant to define **big data as involving a process of analysis that characterises the data involved as big**, rather than as a particular size of product. Thus big data could be seen more as a verb than a noun, and more as a process than an object.

[1] Laney, D. (2001). 3D Data Management: Controlling data volume, variety and velocity (META Group File 949). Retrieved from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

The data discussed in this paper includes, but is not limited to, sources such as social media, mobile phone use, digitally mediated transactions, the online news media, and administrative records.[2] The data we are interested in here can be classified into three main classes of origin: data that is **provided explicitly** (e.g. social media postings, digital survey responses or volunteered geographical information for open mapping); data that is **observed** (such as transactions taking place online or mobile phone call and location records); and data that is **inferred and derived by algorithms** (this would include people's social network structure, trends relating to behaviour or transactions, and economic data such as inflation trends).

The real-time nature of data is also important to the definition of 'big' because it involves a change in the characteristics of information available to those interested in the dynamics of social change. The use of new digital technologies such as mobile phones and internet-based search, communications and transactions mean that in the fields of economic and development we are seeing an emerging process where census-based knowledge, where information is collected every ten years, is being supplemented by constantly updating data which is seldom labelled as 'development'-related, and which may be difficult to access as it flows through corporate, rather than public-sector, circuits of information.

Addressing big data as a process means that we are also interested in the activities related to accessing, cleaning, processing and storing it. As one participant observed: 'we first have to tackle Tedious Data before we can use Big Data'. If we want to avoid problems of reliability and replicability of analysis we have to understand how each relates to an overall objective of promoting positive change and – as importantly – behaving ethically as researchers and activists with regard to data subjects.

3. How is big data being used towards positive social change?

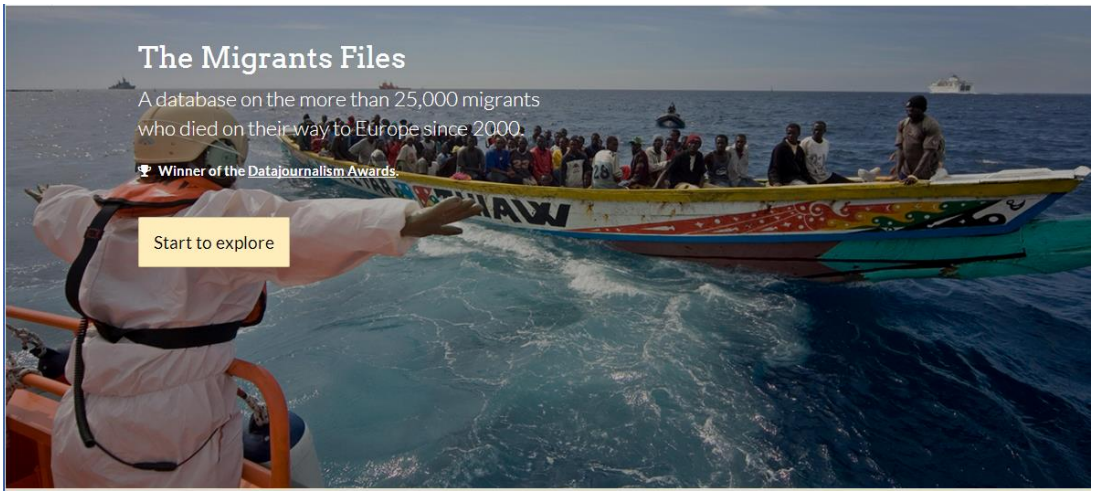
i. Advocating and facilitating

Big data can offer a **new depth of detail** on particular issues. This level of detail can help with advocacy because it makes it possible to show the granular detail of a given problem, and it can be used to create interactive tools which engage the reader and lead them to seek to understand the problem better.

[2] Our definition is similar to that of Kitchin (2014): 'Big Data, new epistemologies and paradigm shifts.' *Big Data and Society*. http://bds.sagepub.com/content/1/1/2053951714528481.full.pdf+html

Box 1. Collecting & visualising information for greater impact

www.detective.io is a service that helps advocates create interfaces for data projects (see http://www.detective.io/detective/the-migrants-files), a project on the human cost of undocumented migration.



The swiftly increasing use of digital communications such as mobile phones and internet connections is also creating opportunities for **activism using volunteered data** in LMICs. Advocates are gaining the ability to aggregate data submitted by individuals more effectively than before, and present it in new ways that can motivate people to action using social media or dedicated platforms such as Ushahidi (www.ushahidi.com).

One example of data activism using volunteered data is the **Black Monday movement** in Uganda (https://www.facebook.com/BlackMondayMovement), begun by civil society organisations in 2012 to publicise stories of corruption and to show that there is money to provide services for people. The activism involved in Black Monday is both locally rooted, publicising local problems and protests, and global in the technology it uses (websites and social media tools). Using Facebook and a variety of other platforms, the organisers of Black Monday have created an online public space for the debate and criticism of powerful interests, which has gone beyond its original goal of exposing corruption to become part of the country's political arena.

Along with this new power to aggregate and expose goes the concern of **too much exposure**. Another facet of data advocacy deals with people's own data profiles, and providing the tools for organisations and activists to protect themselves and the people they are advocating for from unwanted digital surveillance or the inappropriate use of their advocacy tools and information. This makes generating **data awareness and data sovereignty** an important activity in parallel with, or as part of, other advocacy efforts using data.

Box 2. GDELT: tracing media worldwide

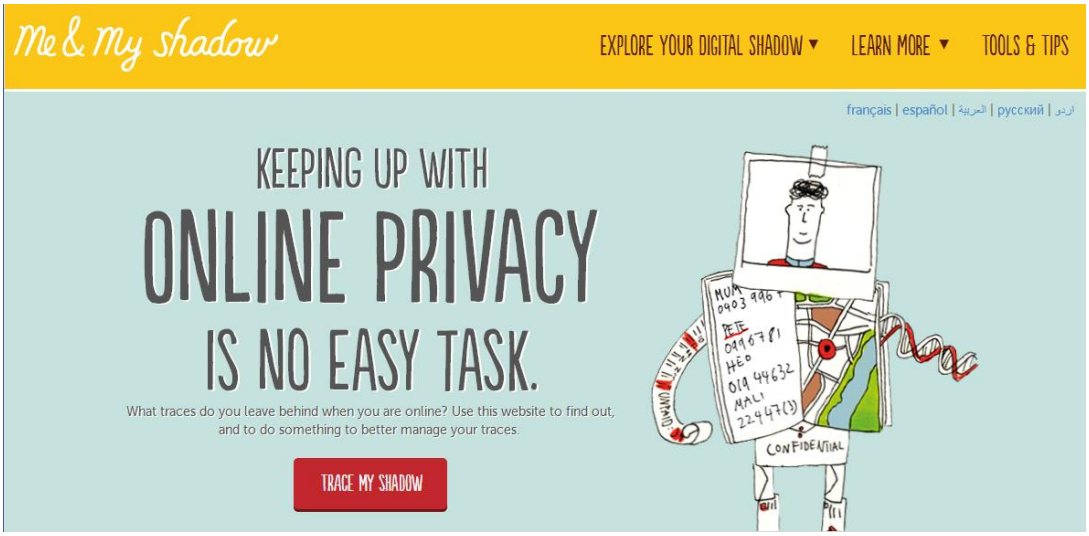
Another project that merges, links and visualises data on worldwide events is GDELT (www.gdeltproject.org), a project that is curated differently from The Migrant Files in that rather than focusing on a single issue, it aims to provide a picture of developing issues worldwide. GDELT brings together real-time broadcast, web and print information from sources worldwide and processes them using natural-language and data mining algorithms to look for events, networks, themes and sentiments present in the information people are posting on the web.

Example: Tactical Technology Collective

The Tactical Technology Collective works to improve activists' and political actors' use of information. They create toolkits and resource guides, films, trainings and events, and serve as an information resource and strategic hub for users of digital information in advocacy. Tactical Tech works in three areas broadly: building digital security and privacy awareness and skills among human rights defenders, independent journalists, anti-corruption advocates and activists; and on strengthening and critically examining pedagogies of security training; second, enabling rights advocates use information strategically and creatively in advocacy, and doing it themselves through projects like Exposing the Invisible – a project about digital investigations; third, a new area of research and capacity building around the politics and practices of using data for advocacy. Tactical Tech also runs **Tactical Studios**, a fee-for-service creative agency for advocacy groups and think tanks developing information visualisations and campaign ideas.

Box 3. Data sovereignty: Tactical Technology's 'Me and My Shadow' project (myshadow.org)

'My Shadow' is a website for a project that raises awareness of how data on what we do online is gathered and used, and what – if anything – we can do to mitigate this. Whenever we use digital devices, surf on the Internet or are active on Facebook or Twitter or any other website, we voluntarily leave information about ourselves. More so, our digital adventures often leave large traces about ourselves without our awareness or complete understanding of what these are. Me & My Shadow has a 'shadow tracer's kit' that helps new users create a visual representation of their own 'digital shadow' based on a checklist of digital activities in their everyday lives.



Building from the success of the My Shadow website, Tactical Tech aims to create and support critical thinking and an increased understanding of data shadows and how they are created, facilitate considered choices when deciding which platforms and technologies to use, and promoting alternatives if possible and advocate for consumer choice outside of what have become 'user and service monopolies'.

Last in this category of data uses, observed big data (i.e. remotely collected data about large groups such as populations or users of a particular technology) can also be seen as a tool for **configuring communities and actions** (sometimes serendipitously) through data science and aggregation. An example is the data science that accompanied US President Obama's 2012 campaign,[3] which categorised and then mobilised voters based on an unprecedentedly detailed level of data on the individual level. Another facet of this phenomenon is the kind of **de facto grouping** that has occurred in the PlayStation 3 community where hackers have pushed the bounds of copyright according to the 'jailbreaking for fair

[3] http://www.technologyreview.com/featuredstory/509026/how-obamas-team-used-big-data-to-rally-voters/

use' decision by US regulators.[4] A community of practice has grown up around the possible changes to the PS3, including features of legal resistance and leadership,[5] which has affected perceptions of copyright, fair use and the rights of technology users.

Problems and challenges for data advocacy

Advocacy using big data raises some particular challenges since it takes place at the interface between the individual and the organisational levels, and between the messy, human level where data is produced and the technical level where it is processed and shaped. Problems identified by advocacy specialists fall into three main categories:

- **Data interpretation**

Organisations and activists must often handle multiple languages (both linguistic and programming) even when dealing with a local phenomenon. When the scale becomes larger (such as with the Migrant Files project or GDELT), the challenge is to find the linguistic and technical capacity to scale up its representation in response.

Language, but also many other issues, cause data to be misinterpreted and lead to false positives or negatives. The metadata (attached information which describes and helps to understand the limitations of a digital dataset) is often missing or incomplete in the case of big data, given that it tends to come from multiple sources and this information may only be available several steps back in the analytical process. Thus the bias of the data used can be unknown, presenting advocates with the choice between using it without fully understanding it, or ignoring it, which may in itself lead to bias. One good example of this problem is social media data, where metadata on who is posting and where they are located may be available, depending on the method used to access the data, but where there is often not enough information to understand the demographic bias of the feed overall[6] – and thus to know whose voices are being heard.

The use of social media data in particular can feature a lack of supplementary information from outside data sources for internal and external validation. (See section 4 on crowdsourcing for practical ways of dealing with these issues.)

- **Organisational challenges**

Organisations in the advocacy sphere have not traditionally had the capacity to do big-data-scale analysis, data storage and computation. Data size and handling issues may therefore be a problem for them, particularly in terms of locating the capacity and willingness to store, compute, utilise, and understand big datasets at organisational level. Organisations need to ask at what point they should choose to invest in this capacity, and understand the indicators that show it will pay off. This paper aims

[4] http://www.wired.com/2010/07/feds-ok-iphone-jailbreaking/
[5] http://www.theguardian.com/technology/gamesblog/2011/jan/13/sony-suing-ps3-hackers
[6] For an example of an analysis of bias in Twitter in the US, see Mislove et al. (2011), 'Understanding the Demographics of Twitter Users':
http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2816/3234

to show ways in which engaging with big data analysis does not have to mean becoming a specialist organisation, but there are aspects of big data analytics which are best done with collaborators working either as consultants (such as Tactical Technology) or on a pro-bono basis, as has occurred with Grameen Foundation's App Lab (see section 3.iii).

Another aspect of the organisational challenge of big data is its often-proprietary nature. Even social media data which appears to be open for use may have legal restrictions (often according to where an organisation is located – for example different rules apply in the EU versus the US in terms of what can be done with data about people), proprietary claims, and particular data science requirements. Related to this, but more broadly, organisations need to consider how to keep sensitive data secure, both because it may be proprietary, but also because even open data may refer to attributes of people which, when read across datasets, may become sensitive.[7]

Finally, organisations looking to do advocacy using big data will inevitably challenge existing power structures and may therefore be subject to political and institutional pressure and /or interference. This is a larger problem that is inherent in advocacy, but the resources and collaborations involved in big data work may lead to particular permutations where pressure can be applied across organisations, or where interference is not immediately visible to the advocates in question.

- **Use and dissemination**

Once they engage in big data analysis, even if through consultants or collaborators, organisations must take responsibility for the data they are using and disseminating. Once data reflecting people's activities, movements, opinions or networks is scraped, downloaded or created by analytics, researchers run the risk of profiling and discriminating – and organisations of allowing this to happen. Thus data in an advocacy context cannot be regarded as neutral, and organisations should consider the risks involved in a given analysis, and its potential impacts on the people in question.[8]

Even though data protection strategies are usually only mandated when organisations deal with data which may be used to identify people, rather than anonymised data, this divide can be deceptive when handling big data due to its distributed nature. When different datasets are made to speak to each other, their combined power may say things about people and communities that would not otherwise be obvious. Organisations therefore need to be aware of cases where data may cement inequalities and detrimental categorisations, and where their work may be feeding into unequal power relations at the point of implementation.

[7] For more on this problem, see Strandburg, K.J. (2014), 'Monitoring, datafication and consent: Legal approaches to privacy in a big data context', in Lane, J., Stodden, V., Bender, S., & Nissenbaum, H. (Eds.). (2014). Privacy, Big Data, and the Public Good: Frameworks for Engagement. Cambridge University Press. (http://goo.gl/iLyLIC)

[8] A useful guide to Privacy Impact Assessments for larger organisations is provided by the UK Information Commissioner's Office:
http://ico.org.uk/for_organisations/data_protection/topic_guides/~/media/documents/library/Data_Protection/Practical_application/pia-code-of-practice-final-draft.pdf

Lessons learned

The points above suggest that organisations dealing with big data inevitably run the risk of technological determinism: Tactical Tech's founders in particular warn of the risks of the **continuing disconnect between technological tools and the strategies that should underpin them**. The tools for dealing with big data are becoming more and more easily available, partly due to the success of community-building (resulting in projects such as the Code for Africa federation, http://www.codeforafrica.org/) – but the analytical work of understanding the implications of the new communities and activities has lagged behind.

Framing questions about strategy and usership needs to happen in parallel with the development of digital tools, otherwise **the political dimension is missing and the tools' power is diminished**. Does it matter that we have access to air pollution information? Not really. It becomes political when someone decides to combine it with other types of information, to map air pollution data over income groups in a city, or housing types, for example. The question is, who will direct the technology to do this? Can it be those with the skills to use the technology, or will this dimension be taken over by other groups and interests?

Privacy and anonymity: the flip side of transparency is **the 'bright light'[9] that digital data shines on users** of reporting platforms and targets of accountability, and that leaves no room for privacy or anonymity. This bright light becomes more risky and complicated for those who are both marginal and 'difficult', i.e. those who are **challenging the state** in some way. So far, the field of digital data for social change lacks coherent conversations about the intersections between platforms, corporations and user privacy, and how these can be managed for the benefit of users. Tactical Tech's new research project, supported by the Making All Voices Count consortium,[10] looks at this flip side by analysing the risks to data and privacy through technology for transparency platforms that area based on, and generate, data.

ii. Describing and predicting

In a separate category from the volunteered data often used in advocacy, **observed data** such as Call Detail Records (CDRs) from mobile phone network operators as well as data generated through the use of social media (known as "social data") are playing an increasingly important role in academic, development and humanitarian research due to its ability to provide high-resolution, dynamic (in terms of time, space and coverage) data sources and methods of analysis. Researchers can combine data which have not previously been combined into higher-resolution datasets from which they can detect new correlations and surface new questions. Mobile phone calling records in particular can provide predictive capacity and real-time information on mobility of people following natural disasters, as shown by the work of Flowminder after the 2010 Haiti earthquake. Similar analyses may be possible to perform in crises, conflicts and elections, although multiple additional sensitivities would need to be addressed in these cases.

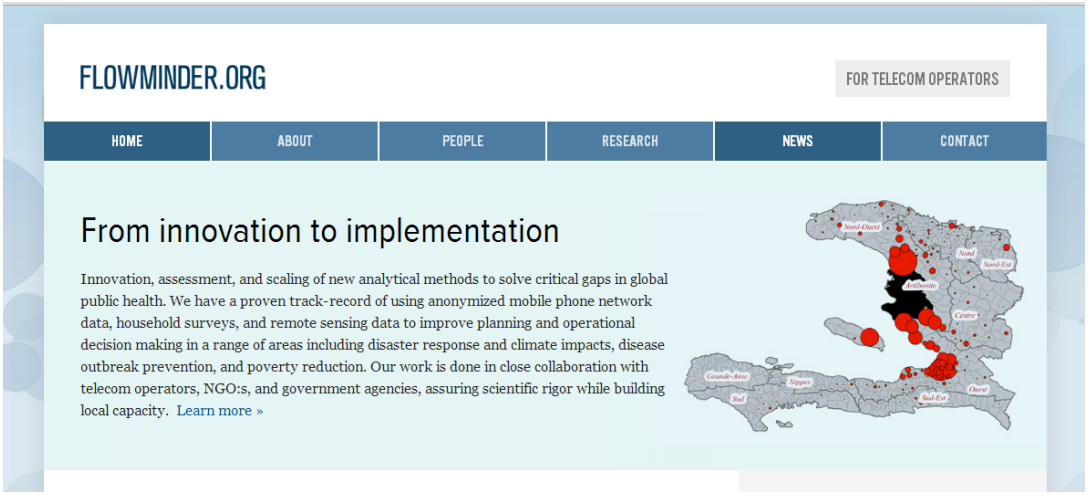
[9] As described by Kate Crawford at the 2014 Transmediale event in Berlin.
[10] http://www.makingallvoicescount.org/

Example: Flowminder

There is an estimated average displacement of over 100,000 people every 2 weeks (1 million displaced on average every second month) related to natural disasters such as floods, earthquakes, hurricanes and monsoons. After the Haitian earthquake in 2010, many people moved away from the capital city, Port Au Prince. Researchers at the Flowminder Foundation asked Digicel, Haiti's biggest mobile phone operator, to share de-identified information about which mobile phone towers subscribers were using when making calls. They distributed reports throughout 2010 to relief agencies to inform about the distribution of displaced populations around the country.

Box 4. Flowminder (www.flowminder.org)

Flowminder's data included the position of 1.9 million subscriber identity modules (SIMs) in Haiti from 42 days before the earthquake to 158 days afterwards, allowing researchers to compare people's movement in the days preceding and following the earthquake. In later research the estimates of geographical distribution of people across Haiti were shown to match well to estimates from a large retrospective UNFPA household survey on post-earthquake migration.



Flowminder's work with the Haiti data brings up several challenges big data researchers must resolve when working in the humanitarian sphere, including access to the data, analytical choices and the verifiability of their conclusions, in order to produce an analysis that is effective and useful in a short space of time (see Bengtsson et al. 2011 for an in-depth account of the project).[11]

Example: Global Pulse

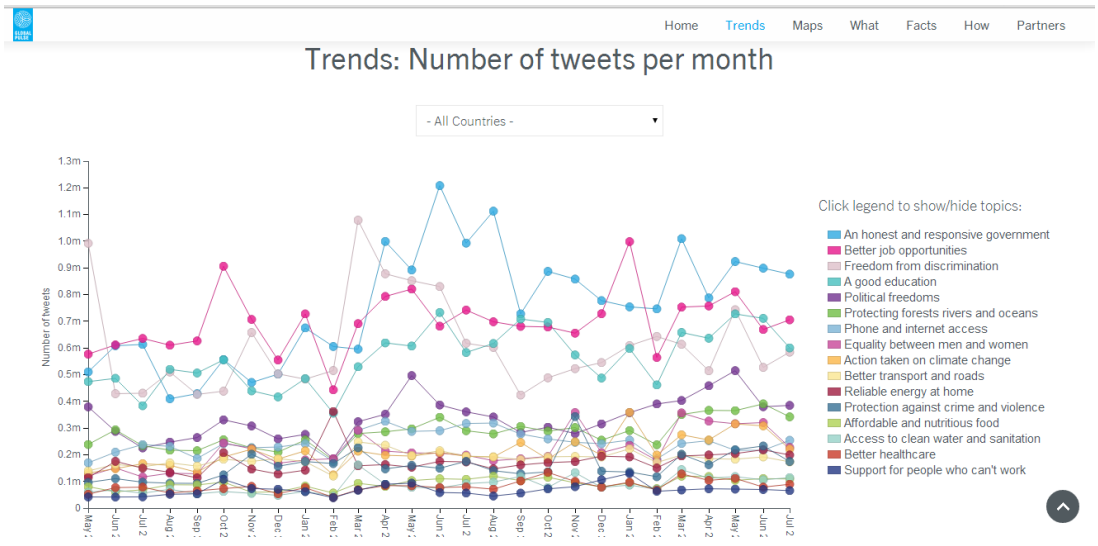
The international development field is reaching the end point for its 15-year effort to achieve the Millennium Development Goals (MDGs). The MDGs constituted a roadmap to halve poverty worldwide

[11] Bengtsson L, Lu X, Thorson A, Garfield R, von Schreeb J. (2011). Improved Response to Disasters and Outbreaks by Tracking Population Movements with Mobile Phone Network Data: A Post-Earthquake Geospatial Study in Haiti. PLoS Med 88: e1001083.

by 2015, providing eight specific targets focusing on poverty, health and the environment. To formulate the next evolution of the MDG challenge, a process of consultation is taking place between the UN, civil society and the private sector. Global Pulse is working to supplement this process with big data and visual analytics, showing how people around the world are debating the topics proposed for the post-2015 agenda online. The interactive visualization shows the 20 countries that have proportionately tweeted the most about each topic.

Box 5. Global Pulse: discussing development on Twitter (http://post2015.unglobalpulse.net/)

Social media data analysis of millions of public tweets identifies relevant keywords, and maps them against Post-2015 priority topics to show which are being talked about the most. The time frame covers 2013 to date and updates every month.



By searching approximately 500 million new posts on Twitter every day for 25,000 keywords (http://datasift.com/essence/kavcgn) relevant to 16 global development topics, the project/dashboard shows which different countries talk the most about a given topic. Filtering in English, French, Spanish and Portuguese yields around 10 million new tweets each month.

Challenges of research with mobile data

- **Access:**

Researchers must first determine the legal process involved in accessing mobile data, which varies from country to country. In some countries a state may have to ask a mobile network provider to make data available to help manage a crisis, whereas in others, states may access communications data for issues such as emergencies crisis and tracking crime. There is some nervousness in the communications industry about how to respond if asked for this data, so researchers may run into significant problems where access is not mandated by law. To deal with this, researchers must build relationships with data

providers ahead of time. If they start the process of gaining access only when a crisis has occurred, there are more 'start-up' challenges.

In the case of Twitter analysis, full access (the 'firehose') can only be achieved through certain intermediaries (DataSift, Gnip), usually at commercial rates. Large non-governmental projects such as those run by UN Global Pulse can make special pro bono agreements with such companies, but replicating this type of research would take similar arrangements or the ability to purchase data. That being said, there are increasing efforts being made by private sector companies to establish grants programs and other formal mechanisms to allow non-profit and social research organizations to gain access to data (see social data analytics company Crimson Hexagon's research grant program,[12] or the recently established Twitter Data Grants program[13]), as well as access to cloud storage and computing which may be needed when handling large data sets for analysis (e.g., Amazon Web Services grants for non-profits - http://aws.amazon.com/grants). Another alternative is utilizing data from the Twitter 'garden hose' which offers 10% of tweets for free to approved research projects.

When it comes to data generated by social media, there is also the question of geographic representation – to understand certain issues, the access problem grows because researchers need to acquire data from other regional and linguistic groups such as Weibo, or the wealth of other microblogging platforms. The boundary between microblogging, blogs and news is also fuzzy in terms of representing and understanding events worldwide, which opens the net even wider in terms of access challenges.

- **Analysis:**

Mobile data is qualitatively incomplete. It offers a way to 'see' population displacement, but can only estimate that people are moving, and not why they are moving or what they need. The data is only available from places where the mobile network is functional, so there may be gaps and biases that the researcher cannot evaluate. For instance, Flowminder's analysis of the Haiti data assumed that people were in the same place where they were last seen. Therefore, if people moved out of one area into another area, but didn't make a call or they moved to an area where the network was down and they were not able to make a call, they were not counted as having moved.

There is also an issue of data source bias. Where the data comes from only one operator or is produced by a particular social group, does that skew the research findings toward a particular demographic? The researcher must consider whether the data covers the entire country or territory in question. Using call records without SMS records included may also skew the data in terms of analysing daily population flows – but is less likely to create bias when following displacement over multiple days and weeks.

- **Validation:**

[12] http://www.crimsonhexagon.com/about-us/newsroom/press-releases/pr-social-research-grant_102513
[13] https://blog.twitter.com/2014/introducing-twitter-data-grants

'Ground truth' is needed to validate researchers' conclusions, but this is seldom available in a crisis-related project. There are several approaches to deal with this problem.

Using mobile data for research:

- First, find out whether the phone users in the dataset are representative of the population, using survey data in combination with the CDR. Once the level of correlation between dataset and population is known, the mobile dataset can then be used as a proxy for population in other models. However, data is only verifiable where a country's census or other survey data is available and accurate. Thus data from lower-income countries with less extensive or accurate census data, or less extensive coverage of phone network infrastructure, can only be validated to a lower standard. For example Bangladesh has 8,000 cellphone antennae compared to Norway's 45,000, making data on users' location less specific; and some LICs such as Myanmar have not conducted a census in decades, meaning that there is no way to accurately validate a CDR.
- Another option is to use multiple sources to validate a CDR: researchers can layer different datasets (some of which may be open data) and use modelling techniques to estimate how accurate each dataset is in comparison to census data.
- It may be possible to add survey questions on top of population data to sharpen mobile data's accuracy around a specific question. Various firms provide a service for surveying mobile users, one being Jana (www.jana.com), where users sign up and are paid to answer questions. These are highly targetable, but fairly costly for the researcher. The advantage of this method is that people give their details when they sign up so the data doesn't have to be conventionally validated; the disadvantage is that they earn a higher rate for answering questions if they are from a frequently-targeted group (e.g. rural women), creating an incentive for people to register inaccurate details.
- In the case of Flowminder's Haiti analysis, the data was initially contentious because it did not agree with the estimates available to responders during the first months. It was later shown to represent both the movement of people and the spread of cholera fairly accurately, however, because it correlated with a later United Nations UNFPA (United Nations Population Fund) household survey on people's self-reported movements.

Twitter analysis:

- Similarly to CDRs, researchers need to know the demographics of social media users in each country of focus. They also need to understand whether there is a difference between self-reported activities or views, and actual actions, and whether Twitter is the typical social media tool.
- In order to verify an analysis, it is also important to understand the differences in opinion between those who use social media and those who do not. To resolve this problem, ideally a follow-on or similar project would test social media 'opinions' against an offline poll or survey results and compare the findings.

Lessons learned

Depending on what is being measured, a **'base layer'** such as a census is often necessary to validate CDR analysis, making big data only as useful as the extent to which a country has old-fashioned demographic statistics.

You can do very different things with the internal corporate data generated by the mobile network operator simply due to the use of mobile phones (such as CDRs) versus surveys conducted using mobile phones, they are differently validated, and may not be comparable.

Mobile network providers are likely to be convinced to share data with a **market-driven approach**: if they can show value from pilot research projects it will make it easier to collaborate with NGOs and data streams will increasingly layer and grow in richness. For these companies, the drive to patent algorithms and to create new features to put more data into models is stronger than the drive to open data. 'we can learn a lot from building new features'.

More **comprehensive ethical systems** need to evolve for the use of detailed personal data such as CDRs: Flowminder is based in Sweden which has a strong ethical assessment system for health data, but ethical committees there are generally unwilling to consider research using datasets from other countries, feeling those should be judged in those countries. So the best system available still falls short with regard to big data, and needs to develop further. Social media presents a different issue – content published online on some social media networks are intended to be and known to the public by the user (e.g. publishing on Twitter, or blogs), while other social networks (like Facebook) give the user a different presumption of how 'public' the content is. ESOMOAR has developed research guidelines[14] for working with social media content for marketing, and other groups like the Big Boulder Initiative[15] are undertaking taking similar efforts to come up with and espouse a common set of social media data usage principles.

The use of mobile phone data analysis has been leveraged by researchers to map population movement patterns in the context of This type of CDR analysis is currently done for natural disasters and to model the spread of disease, but the same techniques could theoretically also be used to measure displacement due to conflict. This brings up further ethical questions: security services and intelligence agencies in certain countries may already have arrangements under which they can access this telecommunications data – such as in a declared emergency or for national security purposes - but should it be accessible for other purposes? Is there a firewall between how the state interfaces with mobile operators from the security versus the humanitarian perspective? With respect to social media data, analytically, representational validity is a major problem, in particular if access is limited to only a subset. In addition, Twitter use differs across geographic regions and countries, and the online dialogue is often steered by a few individuals.

[14] http://www.esomar.org/uploads/public/knowledge-and-standards/codes-and-guidelines/ICCESOMAR_Code_English_.pdf
[15] http://bigboulderconf.com/about

Next steps

Some mobile operators are developing the capacity for CDR research internally (such as Norway's Telenor), but this takes significant resources and is therefore a political process within firms where backing is needed from a high level of management to establish a meaningful research program.

One possible next evolution is for researchers to use mobile phone usage patterns to derive characteristics such as age and gender, and to add a layer of specific survey questions to datasets once they know the base dataset's representativeness. For example, combine 'big data' signals with crowd-sourced data (e.g. SMS-based surveys) to ask for direct feedback on crisis response from affected communities.

iii. Facilitating information exchange

Activities:

- Serving as an information intermediary
- Providing a unified space for collaborative data harvesting from multiple sources
- New interactive platforms for information exchange, aggregation and curation (with enhanced access)
- Information exchange as making new markets and enabling active citizen participation

Potential barriers and obstacles:

- (Meta)data quality and interoperability
- Low interest or buy-in from citizens, governments, other stakeholders
- Potential misuse of exchange mechanism to distribute false data (e.g. a political smear campaign) or to distribute harmful data (e.g. violent or pornography footage distributed)

Big data on human activities stems predominantly from digital communications tools such as Facebook, Twitter and other platforms), and therefore a large part of its utility for objectives of social change is also related to its origins in communication and information exchange. Ways that digital data can be used to facilitate information exchange currently focus around the use of particular platforms to connect citizens and service users with organisations who are representing their needs or conducting interventions. The use of SMS is becoming especially prevalent in developing countries as a mechanism to invite feedback on service delivery, mainly by civil society organisations (CSOs) to support their advocacy strategies, and platforms developed by humanitarian-focused organisations such as Ushahidi focus on SMS in particular as a tool to invite reporting on specific issues. Beyond this, increasingly interactive platforms are also emerging, designed to promote a continuing process of dialogue and feedback between information providers and service users. These platforms (as in the example of Grameen AppLab below) may begin to blur these categories, as the requests and priorities of service users start to shape the flows and types of information provided.

Example: Grameen Foundation's AppLab

Grameen Foundation's work is one example of the way large datasets can be merged and linked around a base layer of data stemming from mobile phones. Grameen Bank's work on microfinance for the poor and unbanked is the largest such enterprise in Asia, and now extends into much of Africa. The related Grameen Foundation has set up an AppLab in Kampala to develop tools which can make use of the information stemming from its clients' microfinance transactions and related information flows. The Bank's target population is those living on less than $2 a day, and working through community knowledge workers (CKWs) it reaches out to clients to form knowledge networks for agricultural information sharing

The survey platform has generated interest from other organisations as a way of surveying Uganda's rural poor, so that now the AppLab works with the World Bank, Barclays, and several other institutions to gather information and evaluate financial products' performance.

Box 6. Grameen AppLab's Community Knowledge Worker program (ckw.grameenfoundation.org)

The AppLab processes and analyses the large datasets stemming from two mobile-based information-sharing applications: 'CKW Search' for information on livestock, crops, and weather; and 'Pulse messaging' – a mobile app which generates popup survey questions and then syncs clients' questions and answers over the web. Through these apps Grameen Foundation is able to check clients' farms' performance and evaluate its agricultural program.

The apps generate large amounts of data: over four years two million searches have been conducted by microfinance clients, each of which is attached to a base dataset on those clients containing a large number of variables on location, poverty level and individual characteristics. Using these linked datasets and ArcGIS or QuantumGIS to do in-depth analysis, the AppLab can track the quantity and type of requests by region and thus identify local problems.

Successes of the AppLab's work so far include the identification of a new crop pest due to an unusual volume of new searches in a particular region. Following up on this early warning signal, Grameen sent out an agricultural extension officer to that region to check, and the pest was contained and dealt with before it could spread.

The survey app, Pulse, has also led to data-driven discoveries such as the realisation that while those involved in active farming are generally women, men are travelling to markets and selling the products. At market time they are often targeted by merchants who encourage them to spend their earnings on consumer products, leading the Foundation to set up a savings product targeted to men at market time, and to market it in the places where clients were selling their products.

Challenges of information-exchange programs

- **The cost of facilitating low-income users' access to technology that enables exchange of information**

National regulations have changed in relation to Grameen AppLab's work. After they had set up the Pulse survey app which syncs regularly to Grameen's network using a pre-paid monthly mobile data package which is free to farmers, the national-level regulation was altered so that a single data package for a month was no longer possible. This problem highlights how researchers and organisations may find themselves responsible for the internet costs or communication charges for SMS from telecoms providers, plus the cost of securing a short code that the people can use for feedback. For smaller organisations in particular these costs can be comparatively high and may discourage certain types of project.

- **Actionable outcomes from data**

Grameen AppLab has found that data quality and interpretation can be problematic, for two reasons. First, because sometimes inputs are bad – survey respondents may answer questions incorrectly. Second, the data is generated under locally specific conditions, reflects clients' engagement with specific processes of farming and finance, and this can make the data difficult to analyse remotely without local understanding.

- **Data storage and processing**

The AppLab has collected four years of data, comprising a dataset that is now too large for conventional analytical tools such as Excel. Collaborating with analytics specialists (in their case, a technology company called Pallantir, which cooperates with them on a pro bono basis) raises questions of who are appropriate 'data handlers', for which new processes and standards have to be developed.

- **Data sharing**

The datasets collected by Grameen Foundation are highly valuable both for development organisations and for those looking to market to Grameen's microfinance clients. Uganda has minimal regulations

regarding customer privacy in communications networks[16], leaving it up to Grameen Foundation to cooperate with mobile network providers around devising solutions for data protection and privacy challenges.

- **Protecting sensitive information**

Grameen Foundation receives regular questions from microfinance clients about privacy (often framed as confidentiality) and data protection. The AppLab's products collect sensitive information such as clients' income level, the size of land they own, their household expenditure, the size of family, and how many children are going to school in their household. In some cases this information is made even more sensitive by clients managing multiple families at once in different locations. Storage and processing issues may arise with sensitive data, particularly where it is backed up or stored with a third party. This problem is set to grow, given that the cloud is often the cheapest place for organisations to store and manage their data, but is also subject to jurisdictional issues if things go wrong and data ownership is challenged.[17]

Lessons learned

Information can multi-task: **multiple feedback loops** are possible within a single program, and apps can be used for data collection on different levels, for different partners. In Grameen's case, via two apps used by farmers via mobile phones, information flows bi-directionally to and from farmers about agricultural problems, services and surveys; to CKWs about their own performance; to other microfinance providers from Grameen's surveys and evaluation; and to the national government from the survey information. There is also an external feedback loop to agriculture experts for more information based on popular searches and new sources.

Multiple flows of information require **multiple forms and levels of data protection and consent**. Grameen Foundation shares none of its data with marketing firms, but does share with collaborating organisations (who get the survey data they fund themselves), approved organisations such as Fair Trade (who get high-level analytics by region), the government's National Agricultural Advisory Services (which receives survey information) and analytics partners such as Pallantir, who get access to anonymised data (removing names, number of children, name of spouse, but leaving unique IDs). In order to use Pallantir as a data analytics collaborator, the Foundation sought permission from funders, and created legal agreements to reflect the permissions involved. On the level of individual clients, Grameen Foundation uses standardised registration-point user agreements involving fingerprint access and a standard consent form. However, additional consent is sought with specific forms whenever a program is added in coordination with an external organisation, so that clients sign an agreement for each additional use of their data.

[16] https://www.privacyinternational.org/reports/uganda/iii-privacy-issues

[17] For an overview of potential problems to do with keeping data in the cloud, see: http://www.e-ir.info/2012/09/14/how-cloud-computing-complicates-the-jurisdiction-of-state-law/

Next steps

Grameen is beginning to think about **predictive uses of the data** it collects, for instance using search patterns to provide information in advance to aid farmers' planning, or making the database more robust with regard to market prices at harvest times.

iv. Accountability and transparency

Activities:

- Enabling participatory contributions to existing functions (e.g. budgeting)
- Drawing connections between existing power brokers and elites
- Exposing information and identifying missing information
- Increasing transparency around data flows
- Checking information
- Slowing down or speeding up digital decay
- New targeted feedback mechanisms
- Individual accountability to facilitate ethical living

Barriers and obstacles:

- Access to technology among low-income users
- Creating sustainability and follow-up
- Risk of detaching data collection and processing from strategy or context
- Defining the value of transparency to motivate participation

Crowdsourcing for accountability and transparency: Chequeado and Ushahidi

Example: Chequeado

Chequeado is an Argentine non-profit independent media outlet that specializes on fact-checking. It is the first initiative of this kind in all of Latin America and one of the four fact-checking organizations -- from over 40 worldwide -- conducting some form of crowdsourcing. For three consecutive years, Chequeado has fact-checked live the presidential speech at the opening of congress, equivalent to the State of the Union Address in the US. In 2014, Chequeado incorporated to this event the use of DatoCHQ, a public database and crowdsourcing platform through which its followers were able to participate in the live fact-checks by sending data via twitter, such as references, sources, facts, articles and questions.

Box 7. Chequeado: online fact-checking in Argentina (http://chequeado.com/datochq)

DatoCHQ was created as a response to the experience of 2013, where over 40,000 individuals followed the event and participated spontaneously sending information using Twitter. Chequeado is continuing to improve this tool and developing other crowdsourcing gadgets, such as its mobile app DatoDuro to receive and send relevant data in real time about ongoing debates, political events and speeches. Chequeado's fact-checking and crowdsourcing method was recently replicated by La Silla Vacía in Colombia to monitor the presidential election of 2014 (http://lasillavacia.com/hilos-tematicos/detector-de-mentiras) and the organization is currently conducting a similar project with El Faro, in El Salvador. There are two other fact-checking initiatives in Chile and Costa Rica, and several throughout the world including in South Africa, Egypt, Ukraine, Italy, in the European Union, among others.

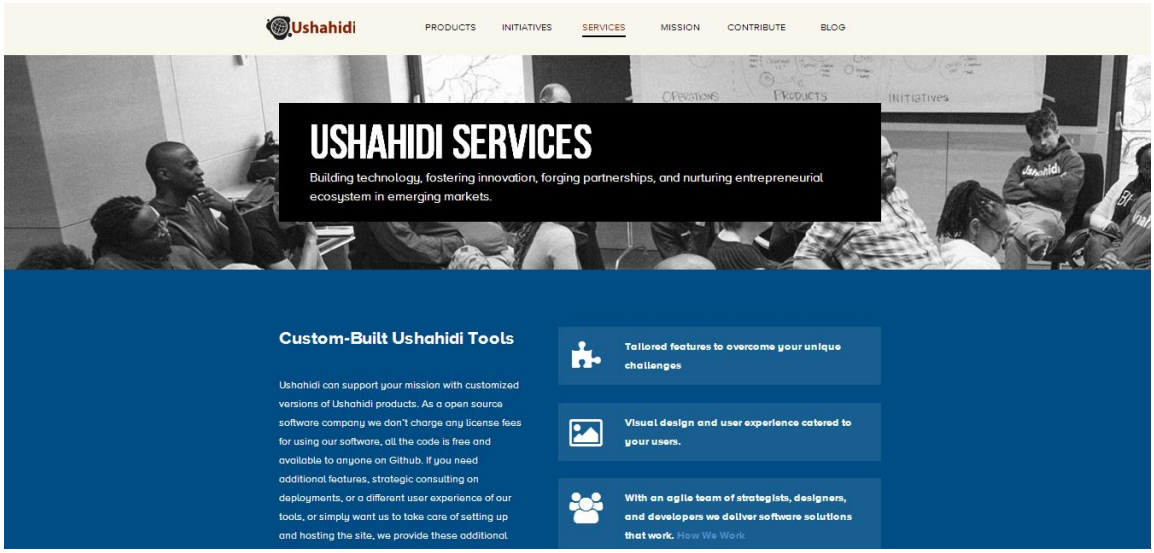
Example: Ushahidi

Ushahidi is a nonprofit technology development company which develops free and open source software for information collection, aggregation and visualisation. They provide users with guidance towards planning around deploying their software through toolkits, based on lessons learned from past deployments, as well as personal experience using it. However, ultimately, it is the deployer's mandate to ensure that they are reaching their target audience and engage with them.

Their work consists of forums, wikis, face-to-face meetings and meetups. They provide an API (Application Programming Interface – a way to access and collect data from websites) with csv (comma-separated values, a common format) and XML (Extensible Markup Language, a newer format for transporting and storing data) download options, and recently released a firehose (crisis.net) that moves data into a similar format for purposes of activism and advocacy.

Box 8. Ushahidi (www.ushahidi.com)

Ushahidi's current deployments include analyses of environmental data; traffic offence reporting; and a 'Stock out' project to provide information about what sort of medicine is being sent to which health centre, and which releases information about how government disseminates medical supplies to different areas and allows people to ask for more information.



Challenges of crowdsourcing

- **Understanding 'the crowd' and how it relates to data ownership**

The work of both Chequeado and Ushahidi raises the question of the ownership of crowdsourced data. What happens if crowdsourcing platforms start selling crowd-generated data and content, as social media companies have done?

In the case of Chequeado's program DatoCHQ, the information is available on a public database available to the public at no cost. Chequeado treats its data as public and tries to make it available to others beyond its own fact-checking activities. Chequeado derives financing from other sources in order not to have to sell data: corporate and individual donors, international cooperation and income generating activities.

Ushahidi's clarified role from direct service provider to (effectively) franchise manager makes it harder for them to know who is using their platform and how. Some new uses, however, favour 'law-abiding people' and may create bias in usership, such as the traffic offence reporting project, and another reporting corruption among police officers. Chequeado deals with this problem in its DatoCHQ project by using a Twitter feed so that users are not anonymous (although identities may belong to an NGO or Think Tank institutional account), and also verifies that information has come from a reliable source before sharing it.

- **Establishing feedback loops**

These are still underdeveloped in the area of crowdsourcing. How can crowdsourcing projects make sure the engaged communities stay tuned in through the project's life-course, and understands what has been done with the information which was generated by people as they participated in the project or platform? This question also applies to the aftermath of events and associated projects. If crowdsourcing methods are used for one-time events, what should happen to the data afterwards?

In the case of Chequeado, 1) data is used in fact-checks, so something concrete and visible to users is produced with the information shared. The data is both processed and fed back to the user who submitted it. 2) Data is made available to the public through DatoCHQ, which as well as being a tool for crowdsourcing operates as a public database. Users can share data (for example a think tank wanting to promote its research) or they can use the database as a source (university students writing papers, or journalists who sometimes struggle to access data). This is possible because Chequeado verifies all the data it makes available. 3) Chequeado also trains other organisations to use data more effectively: its leaders recently coached the editors of the political website La Silla Vacia (www.lasillavacia.com) in Colombia to do live fact-checks on presidential campaign debates. They used social networks, namely Twitter and Facebook, to maintain a conversation with users: users sent in data, suggested items to check, and pointed out when politicians used a false or misleading statement. In this case, social media was employed to generate feedback loops, which in turn, demonstrates the value of crowdsourcing to other organisations.

- **Understanding the data**

How to interpret the data received is a big question in crowdsourcing. Three main issues can be noted: First noise in the data (data which does not add to the information content of a dataset, such as spam, erroneous postings or irrelevant comments) – with passive information gathering, as in the Ushahidi project, where no extra information is available on data sources or context, how are data managers to treat noise?

Relatedly a second issue is verifiability. What are the appropriate mechanisms to deal with (contradictory) crowd knowledge? How can existing approaches be more broadly shared within the community? This becomes more of a question as the 'crowdsourcing community' expands through platforms such as Ushahidi, which offer crowdsourcing technology on a freely available basis with low barriers to entry in terms of skills and capacity. Existing techniques for verification (cf. http://irevolution.net/tag/verification/) may not be known to newcomers or one-time project managers not connected to the broader technology community.

Third, bias in the data. How can data managers know whether the data being submitted is subject to filter-bubble dynamics, i.e. being submitted by a small 'crowd' rather than a large one? And finally, the practical side of the same question: how are project managers to monitor their sources of data, and evaluate 'the crowd's' structure and dynamics?

In the case of Chequeado's DatoCHQ, protecting the quality of the data the organisation shares is very important even in the context of crowdsourcing. The organisation has had many internal debates about how to verify the data and has considered two options: (a) a rating system like Wikipedia's where users

grade each other and check on each other, and (b) having their own team verify the source. They ended up opting for the latter because it was a better guarantee of informational quality, but with the downside is that the process of directly checking each post is very time consuming and taxes the organisation's resources.

Lessons learned

Establish a relationship with the context and the particular crowd involved. Crowdsourcing works where it addresses a **clear demand**, e.g. fact-checking is most appreciated in a context where politicians enjoy little trust. Be transparent, and try to **balance curation and openness**; protect sources and know the rights of whistleblowers so you can seek protection; conduct outreach and integrate your work into other outlets (newspapers, local radio, etc.). **Evaluate the crowd**: who are they, what makes them contribute?

Find ways to **shape and moderate the crowd**: centre the project around a fixed team supplemented by volunteers where necessary; shrink the crowd if needed, since some goals may be achieved through a very small crowd

Verify by **triangulating sources** (never use a single source); by adequately filtering information and checking quality before forwarding posts or sources to all users

Stay independent. Work to diversify the project's or organisation's revenue streams so that it does not become beholden to special interests; take care to sell the project or organisation's overall mission, in order not to become tied to a particular business model conceived to attract potential clients.

Pay attention to issues beyond the technology. Ushahidi's experience shows that **the technology development process is only around 7% of the work**, while the rest is outreach, administration and other 'analogue' tasks. Reporting and compliance can be time intensive and requires dedicated staff. Define success parameters and measures for each sub-project (continuity of the project and/or the amount of data collected).

Next steps

Ushahidi is currently working on a complete rebuild of the Ushahidi platform, dubbed Ushahidi v3, which is set to handle some of the technological challenges such as workflow and verification of data, as well as security. Beyond that, a heavier focus on building effective toolkits is also under way. This will help to guide users of the platform and improve feedback loops.

Ushahidi also continues to reach out to academia and the wider research community on discussions around data sharing ethics, and how to structure their tools to fit around these concerns raised.

Work on CrisisNET (a global firehose of crisis data - http://crisis.net/) also continues, in a bid to make crisis data easily accessible to the wider community in a standardized format.

For Chequeado, the next steps are to explore and expand on how the organisation uses social networks: they can help it appeal to audiences, engage in conversations, and source feedback to show whether opening data can have an impact.

Chequeado will also work to move from engaging only in transparency (the opening of data) to accountability and action.

Finally, the organisation plans to work on strengthening mechanisms for ensuring the veracity of data and communicating that process to its audience, for maximum transparency.

4. Public and private data: open vs. closed?

In thinking about how big data can be used to inform and influence social change, the functional issues of data rights and access, infrastructures and flows are central. These are often considered along the lines of open vs. restricted data sources, which in turn are considered to be aligned with public vs. private data. However, the data sources and uses outlined above suggest that **the boundary between open and commercial data is not clear**, and data's potential uses need not always be divided along the lines of commercial vs. public. In reality there is no single, ideal form of 'open data' and our discussion at Bellagio leaned more toward the term 'opening data', which emphasises the aspects of process and ongoing aspiration – and which may apply to commercially generated data as much as public-sector datasets.

Open data: outlining the issues

There are many forms of data which might be characterised as 'open'. These include:

- Data which has been made freely available by governments, either in the form of tabulated or GIS data which can be manipulated or, as a minimum, data in a fixed format such as PDFs;
- Data which has been rendered legible to the public – or a literate or numerate section of the public – through the use of analysis or curation;
- Data which allows for discussion about its interpretation via the transparent publication of methods used on it, and the opportunity for others to apply different methods in different ways to (potentially) generate different findings

Broadly, if data can be seen as a capacity rather than an object (e.g. it can be used towards accountability, it can lead to the creation of new tools to scrape, access and use it), then open data can be seen as a system of governance and a knowledge commons. Big data may involve some aspects of this, but does not by its nature involve the idea of the commons. The complicated middle ground is important in thinking about this knowledge commons: open data and free data are not necessarily the same. Twitter is open but not free, whereas survey statistics such as the Demographic and Health Surveys are often free but not entirely open.

For open data infrastructures to evolve, it is necessary to resolve the gaps that exist and create a **broader ecosystem of connected actors**, rather than only technologists. In Kenya, for example, the government has been the main actor in the open data movement, but the movement has not been institutionalised and exists only on political will. There needs to be outreach to connect technologists, government, private sector, civil society and academia. In contrast, Uganda had its first open data

meeting three months ago and at the end the national IT authority (civil society organisation) claimed the project because they had the infrastructure and funding to connect government ministries, and could push for policy development.

A provocation: is it a given that 'open data' in a developing country context means making publishing data available online or otherwise making it available in digital form? Given that most people will only ever be consumers of data rather than creating or deeply engaging with data, and that technology access is not evolving equally, it may be worth **broadening the dialogue on open data in developing countries** to include other forms of publicising data. For example, nailing a local authority budget to a town hall door every month may be more open than 'open' digital data that's available online. Conversely, the Open Data Network run by the Open Knowledge Foundation, which is researching open data in the global South, is finding that the different definition of open data in developing countries is justifying governments' refusal to open up data – publishing it in newspapers may not be enough in terms of access. Does this mean that 'open data' is about availability or access? Committing to open data does not equate to committing to making the political process involved transparent, so that in some countries 'data' and 'knowledge' may be in conflict and 'open data' may only represent a form of self-marketing for governments. Under these conditions using big data as a proxy for open data (e.g. as discussed in the Flowminder example above, section 3.ii) may not serve to create change because governments are structurally and politically not open to the possibility.

Private data sources: closed, or not 'Open'?

Sources of data which sometimes occupy a boundary space between closed and open include mobile phone datasets (see Flowminder case, section 3.ii), digital service transaction records, credit cards and financial records. One example of the way that 'closed' commercial data can be opened is through innovation challenges, such as the Orange Data for Development Challenge, in which mobile network provider Orange has opened up limited access to mobile phone datasets generated by its subscriber base in Côte d'Ivoire and Senegal (see www.d4d.orange.com).

The company anonymised the datasets, then challenged researchers worldwide to analyse it and present their findings. The datasets, in order to be 'open' to the researchers (who had to sign a non-disclosure agreement with Orange), were aggregated and abstracted to a level that prevents the mobile phone records from being personally identifiable. The framing of the challenge as an exercise for researchers somewhat restricted the audience who could access the dataset, thus it was not entirely 'public.'

Questions at the intersection of public and private data

Several issues arise when data is made open, or opened with restrictions, as with the Orange D4D dataset. These revolve around fair and accurate representation of the data subjects and of the research community, and risk assessment with regard to privacy and protection of data subjects.

Questions of representation and meaning

- Is it important for the dataset to be interpreted by people with local knowledge, and if not, how can local context be made available to those doing the analysis?

- How can researchers get access to data that is fully representative of the corporate entities with a share of the market – is there a case for pooling data from different providers to make a data commons that is more representative of the territory?
- How are data proprietors engaging with the power dynamics and inequalities in the research field, and consequently with different perspectives on the data, when they make access to the data available to some and not others, for example by publicising the data challenge to a particular research community?
- How can civil society engage with the private sector on its own terms if data access is skewed towards elites?

Questions of privacy and risk to data subjects

- Rather than applying a set of standards to a particular data release, **is it more effective to apply a risk perspective** and devise a risk filter for uses of the data based on the specifics of the data and of the release context?
- Can risk to data subjects be decreased by including an ethnographic dimension to the data analysis that can contribute to understanding the particular risks involved under local conditions?
- Is it appropriate to use a third-party institution to amalgamate data and design rules, taking Institutional Review Boards (IRBs) used in medical research as a model?
- How can data collection, infrastructure and analysis technology help to reinforce the rules and governance around access to data?

Next steps

There is a case to be made for building a **data commons for private/public data** which can function as a way to aggregate data and make it more representative when it is used in research. This would also have to be subject to database ethnographies,[18] standards and risk assessments, since aggregating, linking and merging data present new privacy risks that are different from those of individual datasets.

There is also a case for international actors to lead on this issue, since they have greater data access and analytical capacity. If they can display the value that comes out of the idea of the data commons, it may make it easier to formulate in reality.

Govern data usage, not the data itself: defining specific uses makes specific risk assessments possible. Create rules around the opening of data that relate to the specific purposes and potential of the data in question. A particular dataset may be only employed for certain uses, for example only for discovery (the process of gaining new insights and knowledge by combining or analysing data in new ways), not for taking action.

[18] For more on database ethnographies, see Schuurman (2008): Database Ethnographies Using Social Science Methodologies to Enhance Data Analysis and Interpretation. Available at: http://www.sfu.ca/gis/schuurman/cv/PDF/Database%20Ethnographies%20Using%20Social%20Science.pdf

Organisations advocating for opening datasets must **admit the limitations of anonymisation**,[19] which is currently being ascribed more power to protect data subjects than it merits in the era of big data (see section 3.i for more on this). Data research organisations can lead a dialogue on how to address the risks of opening and sharing data, can devise codes of practice and conduct, and be proactive in saying that even anonymised data may lead back to subjects, or groups of data subjects – otherwise they risk data breaches which will shut down their projects.

5. Conclusions

One issue we kept returning to throughout our discussion at Bellagio was that big data is not for everyone or every purpose. There are many circumstances where small data or analogue approaches are most important, and **big data does not in itself constitute an answer to any particular question**. As noted in the introduction to this paper, big data might best be described as an analytical process, an approach to a question, or even a way of finding the best question, rather than an object in itself. When it becomes the latter, it easily becomes an aspiration or a claim ('big data analysis is inherently better') rather than a tool that is fit for some uses but not others. We have endeavoured to set out in this report some of the uses of big data that stand out as ways of provoking good questions, and that have the potential to open up new landscapes for civil society groups, researchers and funders. Nevertheless, if the examples here highlight one issue, it is that **data always has politics** and that with big data it is especially important to be strategic, and to be able to justify one's analytical choices.

We have outlined a broad range of uses of big data, from high-profile projects by large institutions to smaller, more ad-hoc projects being built from the ground up in response to the possibilities and challenges raised by new digital technologies. Some operate on both levels at once, such as the work of Ushahidi which focuses on both large-scale data processing (as with its new CrisisNET product) and providing a platform for small-scale users to adapt to local needs. The broad scale of projects provides different lessons for different actors, and suggests that funders can have positive impacts by focusing on any part of the environment we have described here that is relevant to their priorities and scale of operation.

The global emergence of big data as a resource has been marked by varying degrees of elitism, mystery and technological determinism. Big data is often controlled by corporate actors, and may be hard to access for civil society groups and smaller-scale activists. Various discourses have become embedded in public understandings of big data that suggest **mystery and power**: data as oil; data as a force to be tamed; data as financial value. This paper has aimed to go beyond the **mythologisation of big data**[20] to show how it is actually composed of small data, which are gathered using methods that have a history and are not necessarily out of reach to smaller-scale organisations; and how 'the algorithm' and 'big data analytics' are not necessarily barriers to understanding and using new sources of digital data or

[19] For more on this, see Ohm, P. (2009). 'Broken promises of privacy: responding to the surprising failure of anonymisation'. Available online at: http://heinonline.org/HOL/Page?handle=hein.journals/uclalr57&div=48&g_sent=1&collection=journals#1713

[20] For more on the metaphors of big data, see Puschmann and Burgess (2014): http://eprints.qut.edu.au/73094/

new types of dataset. We have assessed how different levels and types of data analysis can intersect and inform each other, the tools are being developed that can demystify big data analytical methods, and how **international and national-level projects may cross-fertilise and inspire each other**, as with the example of Chequeado.com, which has inspired groups worldwide to use its strategy and learn from its experiences.

Our analysis here makes a strong case that **it is time for civil society groups in particular to become part of the conversation about the power of data**. These groups, broadly defined, are essential to building **sustainable and relevant big data capacity** on the national level, since they are the connectors between individuals and the level of government, corporations and governance institutions. For big data analysis to be locally driven and rooted, technical expertise and understanding must be built locally and nationally, not only on the international scale in elite data science centres. Civil society groups are also crucially important but currently underrepresented in debates about **privacy and the rights of technology users**. Although big data analytics are a worldwide phenomenon, most LMICs still lack locally enforceable data protection rules and standards. If civil society groups are not involved in exerting pressure for fair principles to guard citizens' data effectively, then the rules and standards will evolve to benefit corporations and governments, leaving the citizen out entirely. Finally, civil society as a whole has a responsibility for building **critical awareness** of the ways big data is being used to sort, categorise and intervene in LMICs by corporations, governments and other actors.

Civil society, research and independent funding has been a force for change in all the most important debates of the past: civil rights, issues of gender and empowerment, health and education among other issues have been placed on the agenda through the efforts of national activists, researchers and funders. Big data is shaping up to be one of the key battlefields of our era, incorporating many of the issues civil society activists worldwide have been working on for decades. It is inevitable that those seeking to promote positive social change will engage with big data as an aspect of power and politics: we hope that the information provided here may help organisations and individuals assess where their particular interests may gain traction in the debate, and what their potential contribution may look like.

To finish, we offer six specific obstacles to the use of big data by civil society organisations in particular which came out of the discussion at Bellagio, with possible approaches to solving them.

Six obstacles to using big data toward positive social change – and potential responses

i. Lack of general data literacy, both top-down and bottom-up

- Different publics: policy, activists, individuals; different actions towards them: upscaling (building capacity internally), upstreaming (collaborating), mainstreaming (wider public)
- Different strategies: gamification/MOOCs/incentivising people by drawing picture of negative consequences/translation of what's at stake
- Think about other ways to achieve communication that don't require tech or don't require literacy or to people who speak different languages.
- Interdisciplinary collaboration

- Disseminate projects to promote data literacy and online 'how to' tutorials
- Include disclaimers in data journalism and visualisation including educating about misrepresentation of data (e.g. http://flowingdata.com/2014/04/04/fox-news-bar-chart-gets-it-wrong/)
- Teach students quantitative skills as part of currently qualitative disciplines, e.g. development studies; teach how to read data with greater sensitivity to bias and methodology; teach qualitative skills to students in data science
- Repository on best practices for addressing different types of data on the part of activists

ii. Lack of open learning environments and repositories

- Broaden the reach of existing spaces such as Coursera and data learning
- Create sandboxes – scout out and raise awareness of experimental open source documentation sites which are exportable, public, secure, e.g. Piratepad, Storify, Titanpad

iii. Lack of resources, capacity and access

- Resources: diversify funding models so funders can't set the agenda for organisations; consortium-based funding rather than institution-based funding
- Capacity: build upon existing structures and publicise collaborative learning environments e.g. consortia of universities or NGOs. Include attention to institutional capacity and organisational change to respond to new data sources (e.g. disaster responders' use of crisis data)
- Create programs and environments that can bring together data science capacity with domain-trained people and area-specific knowledge
- Fund networks rather than individual projects, and establish standards to encourage open tool development (e.g. Knight News Challenge[21] has developed a micro-collaboration tool, [see http://invstg8.net/] to get journalists working in restricted environments access to data more quickly). Cascade responsibility for ethics and standards through networks.
- Encourage experiments in funding schemes. The space is changing too fast to require fixed R&D trajectories up-front. Instead smaller grants with follow-through mechanisms work better to encourage new, untested ideas to take hold. Such schemes also stimulate smaller outfits from the Global South and prevent major players from monopolising funding.
- Encourage crowdfunding initiatives to broaden awareness and increase public support. For example, major funding initiatives could apply matching grants for crowdfunding schemes so that big data projects have a certain buy-in from affected populations (also relates to comments on feedback loops with affected populations)

iv. Challenges of sensitivity and risk perception with regard to using data

- Enlist higher-level authorities on the model of IRBs to advise and filter uses of data,
- Use templates for evaluating data sources e.g. Rebecca McKinnon's Ranking Digital Rights project (rankingdigitalrights.org).

[21] For other products and proposals, see: https://www.newschallenge.org/search?text=encryption&search=all&limit=sincelaunch

- Explore new formats for educating people about privacy/data protection risks (e.g. of Thai medical students having to learn about the history of the person whose body they use for anatomical research)
- Privacy Impact Assessment – checklists for use of datasets. Currently used by privacy officers in corporations and UNGP. But these are very general and give rise to compliance mentality
- Data security:
    - Ethics, storage/infrastructure, regulation dimensions
    - Templates for user agreements for NGOs to standardise behaviours
    - general 'do no harm agreements' before data use, use becomes governable under existing frameworks of criminal law (however this would not eradicate institutional corruption or dangers of criminalisation)
- Guarding against profiling and discrimination
    - Raise awareness of lack of capacity in governance to regulate processes which lead to profiling and discrimination – emerging set of harms we don't yet know how to govern.
    - Use models which have worked in other sectors: IRBs in academia, ombudsman function with access and mandate to respond to malfunctioning algorithms
    - On policy level: make intended impacts of data use clear to data subjects rather than classic user agreements which offer blanket consent – pushes lens of governance away from collection to usage
    - Leadership within the tech community about how to think critically about existing standards which may not be adequate or contextual (e.g. 'proper anonymisation'). Sponsor courses, ambassadors, publicly showcase leaders on data responsibility (funders, policymakers, academia, bloggers). Articulate the problem in new ways
    - Certification scheme (e.g. EU-level) for ethical data research
    - Move to risk-based approach to privacy and data protection rather than standardised rules which are not iterative and flexible enough

v. Storage and computing capacity (also to address the problem of replication)

- Common computing resources e.g. Amazon Cloud, SurfSara (Netherlands); requires dissemination/training on how to use these resources
- Build capacities for independent data repositories

vi. The challenge of externally validating data sources for comparison and verification of big data

- Connect open data movements to the research community
- Create database ethnographies to better understand the data's origins and history
- Take advantage of opportunities for collaborative data cleaning (e.g. data laundry approach), possibly using hackathons. (e.g. Open Street Map)
- Publicise repositories of verification options and techniques

Appendix 1: Conference Participants

Gilbert Byarugaba Agaba
Grameen Foundation, Uganda
Francis Akindès
Alassane Ouattara University, Côte d'Ivoire
Linus Bengtsson
Flowminder, Sweden
Josh Cowls
Oxford University, UK
Maya Indira Ganesh
Tactical Technology Collective, India
Nimi Hoffman
Rhodes University, South Africa
William Hoffman
World Economic Forum, USA
Laura Mann
Leiden Africa Studies Centre, NL
Ulrich Mans
Centre for Innovation, Leiden University, NL
Fran Meissner
European University Institute, Italy
Eric Meyer
Oxford University, UK
Leonida Mutuku
iHub, Kenya
Sophie Nampewo
ACODE, Uganda
Angela Oduor
Ushahidi, Kenya
Carly Nyst
Privacy International, UK
Karin Pfeffer
University of Amsterdam, Netherlands
Ralph Schroeder
Oxford University, UK
Nishant Shah
Centre for Internet and Society, India
Pål Sundsøy
Telenor, Norway
Anoush Tatevossian
Global Pulse, United Nations
Linnet Taylor
University of Amsterdam, Netherlands
Laura Zommer
Chequeado.com, Argentina